Progress Report for NASA Intelligent Systems Program (May 13, 2002)

Automated Knowledge Discovery from Simulators

- PI:
  - Dr. Dennis DeCoste, JPL
- Co-I’s:
  - Dr. William Merline, Southwest Research Institute (SwRI)
  - Dr. Joerg-Micha Jahn, SwRI

Goal and Technical Objectives

- context: simulators play fundamental and growing role in science and engineering investigations
  - across NASA, DoE, DoD, FAA, industry, and academia
  - other related research on simulation focuses on high-performance computing issues (i.e. speed and accuracy of *each* simulation run)
- **our goal**: *develop data mining methods that enable scientists to exploit unique nature of "science by numeric simulation"*
  - couple leaders in machine learning (JPL) and science simulation (SwRI)
  - unique abilities of simulators include:
    - can generate vast, potentially unlimited, volumes of new data
    - explores conditions unlikely/impossible from state-of-art universe observation
  - fundamental technical challenge in harnessing simulators for science:
    - automatically determine focused *set* of simulations runs to do, so that:
      - *maximize science throughput* (e.g. max "new knowledge per simulation week")
      - overcome infeasibility of standard "uniform sampling" of simulation space (which scales exponentially in the number of parameters being studied)

Technical Problem Statement

- automatically determine set of simulations to run
  - goal: discover parameter values (initial conditions, dynamic variables, etc.) leading to behaviors of scientific interest
    - **concisely**: learn predictive models (vs memorizing tried value sets)
    - **efficiently**: via intelligent (vs uniform) sampling of parameter space
  - one of our example "science by simulation" target problems:
    - "identify boundary conditions for when binary asteriod pairs form"
      - 3 parameters: impactor velocity & angle and impactor/target mass ratio
    - traditional uniform sampling is prohibitively wasteful
      - e.g. 1000 runs for just 10 impact velocities, 10 angles, 10 mass ratios
      - unsuitable in general as well, for any realistic/complex science:
        - » simulation costs necessarily scale exponentially with the number of parameters, regardless of importance of each parameter
        - » at 1 simulation/week, requires 20 machine years, for this *simple* example

Technical Approach

- our solution involves several key components:
  - event detectors – define scientific behavior(s) of interest
    - determine whether/where event occurs in a simulation trace
    - e.g whether binary asteriod eventully occurred after impact
    - to amortize savings, explore multiple events during same sims
  - “active learning” over simulator’s parameter space
    - given simulations so far, determines which parameter settings to use for <u>next</u> new simulation run(s), to best improve current predictive model of the conditions leading to behavior of interest
  - state-of-the-art support vector machine (SVM) classifiers
    - SVMs are core machine learning technology we use, because:
      - learns robust non-linear predictive classifier models
      - solid basis for active learning (via rapid “version space” reduction)

Data and NASA Relevance

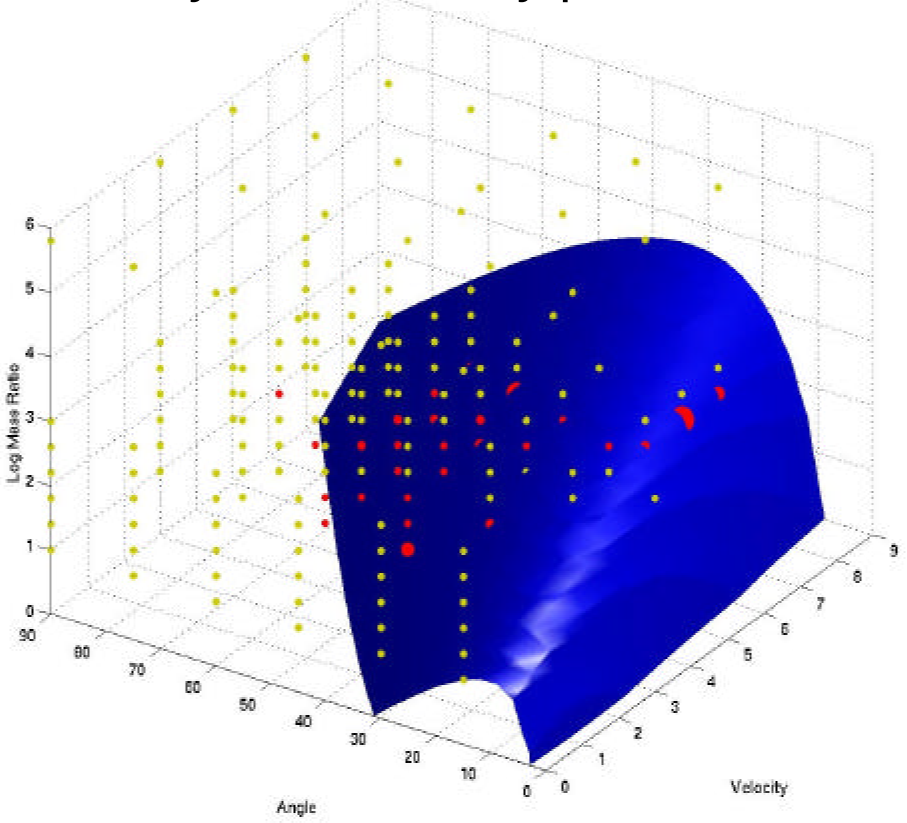
- initially focus experiments on 2 NASA science simulators:
  - Asteriod Impact simulator
    - lead scientist: Dr. William J. Merline, SwRI
    - relevance: formation of asteriod satellites is not yet well understood and is difficult to study without simulators, since they are rare and difficult to find (e.g. via current telescopes). Better understanding the preconditions for such behaviors can lead to guidance in where to look for physical confirmation of scientific theories. Finding and understanding satellite formation and evolution is fundamentally important to space science. For example, presence of satellites is sole means of understanding object density, in lieu of spacecraft flyby or extremely rare detectable perturbation on nearby planets. Discovery of Dactyl binary asteriod and followup discoveries have sparked ongoing revolution in asteriod astronomy.
    - Magnetospheric Dynamics simulator
    - lead scientist: Dr. Joerg-Micha Jahn, SwRI
    - relevance: understanding the dynamics of the magnetosphere is both of fundamental science interest as well as vital for prediction of “space weather” and its disruptive consequences (e.g. hazards to communication satellites, power grids, spacecraft, and astronauts).

Accomplishments & Preliminary Findings

- one key technical innovation to date was:
  - radically increasing traditional SVM classification speed (10-100x)
    - this advance in the core SVM technology was required to enable us to scan over enormous candidate parameter spaces and select the settings for the next simulation run which are likely to most improve the current predictive model (which maps from parameter states to expected simulation outcome, e.g. binary asteriod or not).
    - other accomplishments include:
  - developing an initial prototype of the basic active learning software
  - obtaining preliminary results on our “asteriod satellites” simulator problem [see next slide for example result graphics]
    - these preliminary results include uniform sampling results, for comparison to the intelligent sampling approach

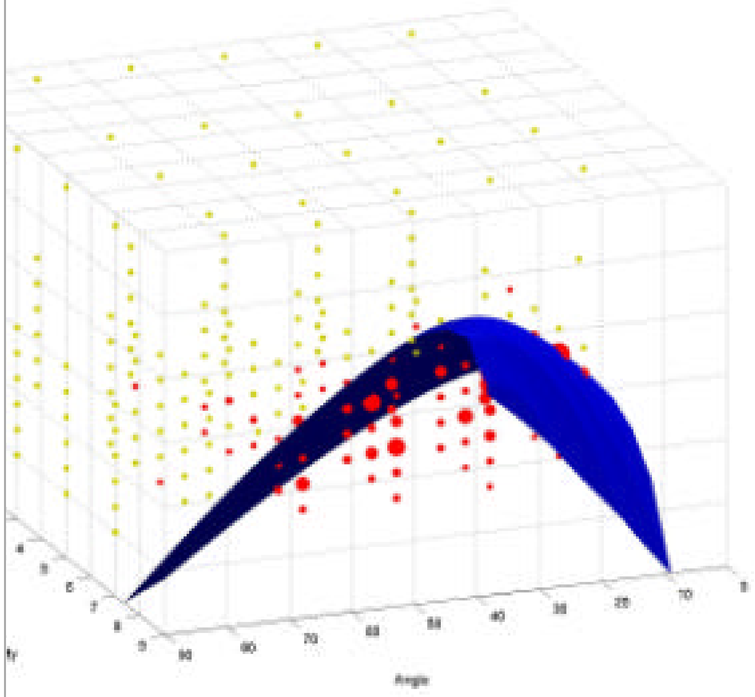
Accomplishments & Preliminary Findings (cont)

- example results for Asteriod Impact simulations
  - red/yellow = binary pair occurs/not; blue = prediction discriminant



3d view

(showing learned SVM quadratic discriminate surface for predicting whether binary pair will occur for given parameter values)



second 3d view

(showing that most strong pair occurances occur on under-side of discriminate boundary)

Technical Significance of Progress / Expected Impact on NASA

- our initial results with the Asteriod Impact simulator illustrate the promise of our basic approach, for radically reducing the number of simulations required (relative to uniform sampling).
- impact: maximize “bang for buck” for “science by simulation”.
- our SVM innovation, giving orders of magnitude speedup of classification, is likely to have wide-spread impact on our machine learning field as well as many NASA applications.
- impact 1: makes SVMs competitive/superior speed-wise with popular alternatives (e.g. neural networks) for which SVMs have already been demonstrated to often be superior otherwise (accuracy, robustness).
- impact 2: makes SVMs practical in new applications (e.g. real-time classification onboard resource-constrained spacecraft

Technical Significance of Progress / Expected Impact on NASA (cont)

- we expect huge and varied payoffs of this project:
  - simulation runs to date have *already* lead to new insights:
    - high-speed collisions having higher prevalence for making moons
      - our intelligence sampling approach naturally exploits such discoveries, focusing more simulation effort on exploring (and confirming) them.
      - our simulations are suggesting that small asteriods with small moons seem more prevalent than expected
      - would be difficult/costly to see with telescopes, due to small sizes
      - after more careful analysis (using intelligent simulation sampling), scientists will have strong justifications for why/whether such telescope costs are likely to be worthwhile. ***This could majorly impact how future science and observation planning is done throughout NASA, improving science throughput*.**
  - our approach will also provide foundation for other similar innovations
    - e.g. including extensions to instrument design and observation planning

URLS Describing Team

- PI’s page:
  - http://www-aig.jpl.nasa.gov/home/decoste/dmd-pubs.html
- Co-I’s page:
  - http://www.boulder.swri.edu/~merline
  - example images of actual asteriod satellites discovered
    - http://www.boulder.swri.edu/~merline/decoste
- online movies of Asteriod Impact simulation runs:
  - http://www.boulder.swri.edu/~benke/present/sims/results_movies.html

Facilities Used / Personnel

- JPL
  - Dr. Dennis DeCoste (PI), machine learning
    - Dominic Mazzoni, computer scientist
  - 100-node Linux Beowulf machine, used for simulations
- Southwest Research Institute:
  - Dr. William Merline (co-I), planetary scientist
    - Brian Enke, computer scientist
    - also: Dr. Dan Durda and Dr. Bill Bottke, consultants on satellite formation modeling
  - Dr. Joerg-Micha Jahn, space physicist
    - Anders Johanson, space engineering

References

- Papers
  - D. DeCoste. **Anytime Interval-Valued Outputs for Kernel Machines: Fast Support Vector Machine Classification via Distance Geometry**. *Proceedings International Conference on Machine Learning* (ICML-02), July 2002.
  - E. Mjolsness and D. DeCoste. **Machine Learning for Science: State of the Art and Future Prospects**. *Science*, Volume 293, pp. 2051-2055, September 14 2001.
- Presentations
  - Invited tutorial, **Support Vector Machines and Other Kernel Methods: Key Concepts, Recent Advances, and Applications**, Institute for Pure and Applied Mathematics (IPAM), Conference on Mathematical Challeges in Scientific Data Mining, UCLA, January 17, 2002.

References (cont)

- papers that asteriods simulations so far have influenced:

  - Merline, W.J., Weidenschilling, S.J., Durda, D.D., Margot, J-L.,Pravec, P., and Storrs, A.D. "Asteroids Do Have Satellites", in Asteroids III (eds. W.F. Bottke, A. Cellino, P. Paolicchi, & R.P. Binzel), Univ. of Arizona Press, in press (2002) [review chapter in major asteroid research/reference book]
  - Merline, W.J. "Progress on the Search for Binary Asteroids", Nature (invited review), in preparation (2002).
  - Merline, W.J., Close, L.M., Dumas, C., Chapman, C.R., Menard, F., Owen, Jr., W.M., Potter, D.E., Slater, D.C. "Using Adaptive Optics on Large Telescopes to Search for Satellites of Asteroids", Proc. SPIE Conf. on Astronomical Telescopes and Instrumentation: Discoveries and Research Prospects from 6-10m Class Telescopes II", in preparation (2002).
  - Merline, W.J., Close, L.M., Siegler, N., Dumas, C., Chapman, C., Rigaut, F., Menard, F., Owen, W.M., and Slater, D.C. "S/2002 (3749) 1", IAU Circ. 7827,2 (2002)